4.5

REGRESSION ANALYSIS IN PLANNING RESEARCH

Helen Bao

1. Introduction

Regression techniques are used to model causal relationship between an outcome variable and one or several of its determining factors. When applied to planning research, it can be a useful tool to study a wide range of topics. A quick survey of some recent publications in key planning journals shows that regression methods have been used to investigate the effectiveness of planning policies (e.g., Chellman & Ellen, 2011; Delang & Lung, 2010; Greasley et al., 2011; and Stagoll et al., 2010), the pattern of urban growth (e.g., Deng et al., 2009; and Joseph & Wang, 2010), residential choice behaviours (e.g., Gao & Asami, 2011; Hoshino, 2011; and Kahn & Morris, 2009), trends in urban designs (e.g., Dumbaugh & Li, 2010; and Ryan & Weber, 2007), and environmental issues (e.g., Drummond, 2010; Lubell, et al., 2009; and Schweitzer & Zhou, 2010), to name a few.

Regression techniques work well when relationships are reasonably well defined (either in theory or based on empirical evidences), and data are of good quality and quantity. From a practical point of view, a deep empirical literature or an existing theoretical framework on the subject matter will help researchers specify their regression models correctly. All else being equal, a large dataset always gives researchers better chances to model planning issues accurately. In the appendix a list of planning publications using regression techniques is given for 2010 and 2011. A notable pattern is that most of these studies have a large number of observations and variables in their final models, regardless of the country and planning issues involved.

The basic relationship modelled using regression methods in planning research can be described using equation (1).

\[
\text{Planning Outcome} = f (\text{Planning Factors, Control Variables}) \tag{1}
\]

More specifically, in equation (1) Planning Outcome is specified as a function of Planning Factors and Control Variables (i.e., Planning Outcome is determined by both Planning Factors and Control Variables). For example, to study the effectiveness of a primary school subsidy programme by local governments, the Planning Outcome variable could be a school performance score. The Planning Factor variable could be an indicator of whether the corresponding school received the
subsidy. Control Variables will include all other important determinants of school performance, such as teacher-pupil ratio. Regression methods can be used to estimate equation (1) in order to determine whether and how planning outcome is affected by planning factors, holding other control variables constant. When used appropriately, it is a powerful tool to isolate the net effect of planning policies.

The popularity of the method roots in its sound theoretical foundation and its flexibility in applications. Hence it is important to understand the technical background of regression methods, the procedures of performing a regression analysis, and some practical issues facing researchers in planning studies. The objective of this chapter is to provide a concise overview of key technical details of regression analysis, and a practical guideline on how to apply regression techniques in planning research. For brevity, this chapter will focus on linear regression analysis only. Other types of regression methods, such as non-linear regression, logistic regression, and panel regression, are extensions of standard linear regression techniques. These advanced topics will be left to the readers to explore when linear regression techniques are inadequate or inappropriate for their analysis.

The rest of the chapter is organized as follows. Section 2 covers the concept of linear regression and the estimation of linear regression models using the ordinary least squares (OLS) method. Some practical issues with regard to the application of linear regression methods in planning research are discussed in Section 3. Section 4 concludes.

2. Linear regression models and ordinary least squares (OLS)

2.1 OLS explained

Linear regression models reveal a causal relationship between a dependent variable \( Y \) and independent variables \( X \). The dependent variable is the effect, and is placed in the left-hand side of the model. Independent variables are the causes, which show up in the right-hand side of the model. In linear regression models the rate of changes of \( Y \) in response to one unit change in \( X \) remains constant (i.e., a linear relationship). For example, a researcher would like to study the relationship between planning policy effectiveness and local government’s fiscal capacity. The following linear regression model was established.

\[
PEI = 20 + 0.04HI
\]  

(2)

where \( PEI \) is a policy effectiveness index that takes a value from 1 to 100, and \( HI \) is the average monthly household income in pounds (serving as an indirect measurement of local fiscal capacity). In this example \( PEI \) is the dependent variable and \( HI \) is the independent variable.

According to the model, when average monthly household income increases by one pound, the policy effectiveness index will increase by 0.04 points. More specifically, when \( HI \) increases from 1,000 pounds to 1,001 pounds or from 1 pound to 2 pounds, the changes in \( PEI \) will be the same. The relationship is illustrated in Figure 4.5.1. Note that the relationship resembles a straight line (i.e., a linear relationship).

If, however, the researcher suspects that in the higher income range \( PEI \) will be less responsive to changes in \( HI \), equation (2) will not be appropriate. Consequently, the first question to ask before establishing a linear regression model is ‘Are the two variable linearly related?’ If the answer is ‘no’, either alternative approaches (non-linear regression, for example) should be adopted, or variables should be transformed. Non-linear regression is beyond the scope of this chapter. Discussions on variable transformations can be found in Section 2.4.
A linear regression model is found by minimizing the average differences between the predicted values by the regression model and the observed value of $Y$. Suppose we want to estimate a regression model of $Y$ using two independent variables $X_1$ and $X_2$. The unknown regression model is

$$Y = \beta_0 + \beta_1X_1 + \beta_2X_2 + \epsilon$$

(3)

where $\beta_0$ is the value of $Y$ when all independent variables equal zero, $\beta_1$ and $\beta_2$ are coefficients to be estimated, and $\epsilon$ accounts for the variability in $Y$ that cannot be explained by the linear relationship between $X$s and $Y$.

Equation (3) can be used to reveal the relationship, if any, between $Y$ and each of the two explanatory variables, and to forecast the value of $Y$ given the value of $X_1$ and $X_2$. $\beta_i$ determine the relationship between $Y$ and $X_i$ ($i = 1$ or $2$ in this example), after the effect of other independent variables is controlled for – for example, if $\beta_1 = 0X_1$ and $Y$ have no linear relationship because $Y$ does not change when $X_1$ changes. If $\beta_1 > 0 (< 0)$, $X_1$ and $Y$ are positively (negatively) related because the two variables change in the same (opposite) direction, net of the effect of $X_2$. In other words, $\beta_1$ captures the ‘net’ effect between $X_1$ and $Y$, because the impact from $X_2$ is reflected in $\beta_2$. If $X_2$ is omitted from (3), the estimated value of $\beta_1$ could be misleading as it may contain the combined effect of $X_1$ and $X_2$.

Based on the foregoing discussion, it is important to include all of the determinants of $Y$ in a linear regression model in order to correctly identify the effect of each included independent variable. Choosing the right set of variables is a challenging task, and some selection algorithms will be helpful. This is discussed in the ‘Variable selection’ section in the later part of this chapter.

We need to find the estimates of $\beta_0$, $\beta_1$, and $\beta_2$ so that the relationship between $Y$ and its determinants can be identified, and a prediction can be made. The estimated model

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2$$
should have the smallest *Sum Squared Residuals* (SSR) as measured by the following formula.

\[
SSR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_{i1} - b_2 X_{i2})^2
\]

The value of \( b_0, b_1, \) and \( b_2 \) can be found by minimizing \( SSR \). The technique is called the ordinary least squares (OLS) method. All statistical software packages have procedures for OLS estimation. The information reported is fairly standard. The *t* test (or *t* test in short) is the statistical tool to detect linear relationships between \( Y \) and each of the \( X \)s. The *null hypothesis* of a *t* test is ‘the corresponding independent variable is not linearly related to \( Y \)’. In a statistical test, the null hypothesis is the ‘default state’ that can be either accepted or rejected. If the *p-value* of a *t* test is smaller than the selected significance level, the null hypothesis should be rejected, and the corresponding variable should be kept in the model.\(^2\)The procedures of *t* test are described in Figure 4.5.2.

The following example gives a standard regression output using Excel, and the procedures to interpret coefficient estimates.

To estimate the implicit value of buildings’ energy performance, the following regression models were estimated using house prices (\( PRICE \)) as the dependent variable, and an energy performance rating (\( EP \)) as the independent variable. In this example \( EP \) takes a value from 0 to 100, with 100 as the best energy performance rating. The model to be estimated is given as equation (4).

\[
PRICE = \beta_0 + \beta_1 EP + \epsilon
\]  

(4)

---

**Figure 4.5.2** Procedures of the Student *t* test.
Regression analysis in planning research

Table 4.5.1 Excel output for equation (4)

<table>
<thead>
<tr>
<th>Regression Statistics</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple R</td>
</tr>
<tr>
<td>R Square</td>
</tr>
<tr>
<td>Adjusted R Square</td>
</tr>
<tr>
<td>Standard Error</td>
</tr>
<tr>
<td>Observations</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ANOVA</th>
</tr>
</thead>
<tbody>
<tr>
<td>df</td>
</tr>
<tr>
<td>-------</td>
</tr>
<tr>
<td>Regression</td>
</tr>
<tr>
<td>Residual</td>
</tr>
<tr>
<td>Total</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Coefficients</th>
<th>Standard Error</th>
<th>t Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-760,044.83</td>
<td>-2.15</td>
<td>0.03</td>
</tr>
<tr>
<td>EP</td>
<td>12,223.27</td>
<td>3.22</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>

Using Excel Data Analysis Add-ins (Regression), the regression model was estimated and the output is given in Table 4.5.1. The value of $\beta_0$ and $\beta_1$ are -760,044.83 and 12,223.27 respectively (in the column labelled ‘Coefficients’). The P-value of the t test for EP is less than 0.01. At the significance level of 5 per cent, the variable passed the t test. Therefore one can conclude that a significant linear relationship between PRICE and EP is identified. When energy performance rating increases by one point house prices will increase by 12,223.27 pounds on average. Note that if the P-value is greater than or equal to 5 per cent in this example, the conclusion would be that EP has no impact on house prices, or homebuyers are not willing to pay extra for houses with better energy performance.

2.2 Model evaluation

Regression models need to be evaluated before putting them into use. Poorly estimated models can provide misleading information for decision makers. The first step in model evaluation is to check the overall model fitting (i.e., whether the model provides a reasonably good estimation and prediction). The following tools are commonly used for this purpose.

1. **R Square**
   R Square measures how much variation in $Y$ can be explained by $X$s using the estimated regression model. Its value is between 0 and 1, with 1 indicating a perfect model, and 0 for a completely useless model. The acceptable range of R Square varies across areas. One should always consult relevant literature and/or experts in the field for appropriate benchmarks.

2. **Adjusted R Square**
   R Square has an inherent shortcoming – it always increases when new variables are added to the regression model, even if those variables do not add any values to the estimation. Hence if R Square is used to compare models with different sets of independent variables, models with
more variables are often selected. But those models are not necessarily the best ones. To solve this problem, Adjusted $R^2$ is introduced by taking into account of the number of independent variables. Adjusted $R^2$ will increase only if the added variables make significant contributions to the model estimation. It is useful when comparing several regression models.

3. $F$ test

The null hypothesis of an $F$ test is that all coefficient estimates equal zero (i.e., none of the independent variables has a linear relationship with $Y$). $F$ test is used to answer the question ‘is the model is completely useless?’ It is equivalent to a test of ‘$R^2$ is zero’. When the $p$-value of the test is smaller than the selected significance level, the aforementioned null hypothesis can be rejected, and one can conclude that the model is useful.

If a model passed its $F$ test, it means that at least one of the included independent variables has a linear relationship with $Y$. But $F$ tests cannot tell which one it is. We need to use the student $t$ test to identify the determinants of $Y$. More often than not, not all variables in the data set are important in determining the value of $Y$. These ‘useless’ variables can cause problems and should be removed from the model. Hence selecting the correct set of independent variables is an important step in regression analysis.

2.3 Variable selection

In practice researchers often have more than one independent variable at their disposal. Appropriate statistical tools and selection algorithms should be adopted to decide the correct set of independent variables to be included. The presence of unimportant (i.e., statistically insignificant) variables or absence of important variables may cause misleading coefficient estimates. The following example illustrates the aforementioned problems.

Using the economic prices of good energy performance rating examples again, the researcher obtained additional variables to construct two more regression models. The specification of the three regression models are given here.

Model 1: $\text{PRICE} = \beta_0 + \beta_1 \text{EP} + \epsilon$

Model 2: $\text{PRICE} = \beta_0 + \beta_1 \text{EP} + \beta_2 \text{SIZE} + \epsilon$

Model 3: $\text{PRICE} = \beta_0 + \beta_1 \text{EP} + \beta_2 \text{SIZE} + \beta_3 \text{NEW} + \beta_4 \text{FAST} + \beta_5 \text{DIST} + \epsilon$

where $\text{SIZE}$ is the floor space of the property in squared metres, $\text{NEW}$ equals one if the property is a new build, $\text{FAST}$ equals one if the property is near a fast train link, and $\text{DIST}$ is the distance to the nearest primary school in metres.

The regression outputs are summarized in Table 4.5.2. In Model 1 the $R^2$ square is merely 0.05, which indicates a poor fit of the data. $\text{EP}$ explained only 5 per cent of the variation in $\text{PRICE}$. One cannot put much faith into this model for either predicting housing prices or for estimating the marginal price of buildings’ energy performance. Once $\text{SIZE}$ is added to the regression model, the regression output is changed significantly. First of all, $R^2$ square is improved by more than 85 per cent. This is not surprising as house size is an important determining factor of house prices. It is reasonable that $\text{SIZE}$ is able to explain about 85 per cent of the variation in $\text{PRICE}$. Secondly, $\text{EP}$ is insignificant at the 5 per cent significance level now (i.e., $P$-value=5%). This is because by omitting $\text{SIZE}$, Model 1 is mis-specified, and the coefficient
estimate is misleading. Model 1 indicates that \( EP \) is an important variable in determining the value of \( PRICE \). However, based on Model 2, which is more reliable given a much higher \( R^2 \), the linear relationship between \( EP \) and \( PRICE \) cannot be established. In Model 3, the researcher added three more independent variables. However, none of these variables passed the \( t \) test at the 5 per cent level. If we use R Squares to compare Model 2 and Model 3, we will select Model 3 because its \( R^2 \) Square is larger. However, the Adjusted \( R^2 \) Square of Model 3 is smaller. This is because some useless variables are added to the model. Variables \( NEW \), \( FAST \), and \( DIST \) cannot pass their \( t \) tests, and hence should not be included in the regression model. This is a good example of problems associated with including unnecessary variables in regression models.

Of course, one should not remove all insignificant variables at once (e.g., remove \( EP \), \( NEW \), \( FAST \), and \( DIST \) in one step). This is because the calculation of \( p-values \) uses information of all variables included in the regression model. If one variable is removed, values of \( p-value \) of the remaining variables will be affected, and some of them may turn out to be significant. However, finding the best model by checking variables one at a time can be a fairly tedious procedure, especially when a large number of independent variables are involved. Almost all statistical software packages have procedures to perform these tasks automatically. Excel, unfortunately, does not have this feature so far. Figure 4.5.3 summarizes the three most commonly used variable selections algorithms. The backward elimination method starts with a full model (i.e., a model with all variables available), and removes one insignificant variable at a time. The forward selection method works in just the opposite way. Both methods adopt a ‘one-way only’ approach, meaning that once a variable is removed (added), it will never be added in (removed from) the model. Consequently the final models resulting from these algorithms may contain insignificant variables (if forward selection is used), or overlook some important factors (if backward elimination is used). The stepwise regression method, on the other hand, addresses these drawbacks by allowing variables to re-enter or exit the model at any stage of the selection process. Consequently this is the most popular method among the three.

### Table 4.5.2  Regression using \( PRICE \) as the dependent variable

<table>
<thead>
<tr>
<th></th>
<th>Model 1</th>
<th>P-value</th>
<th>Model 2</th>
<th>P-value</th>
<th>Model 3</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>(-760,044.80)</td>
<td>0.03</td>
<td>(-259,704.70)</td>
<td>0.02</td>
<td>(-186,727.30)</td>
<td>0.12</td>
</tr>
<tr>
<td>( EP )</td>
<td>12,223.27</td>
<td>&lt;0.01</td>
<td>2,375.48</td>
<td>0.05</td>
<td>1,492.33</td>
<td>0.25</td>
</tr>
<tr>
<td>( SIZE )</td>
<td>—</td>
<td>—</td>
<td>4,762.12</td>
<td>&lt;0.01</td>
<td>4,746.35</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>( NEW )</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>26,545.81</td>
<td>0.20</td>
</tr>
<tr>
<td>( FAST )</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>4,088.92</td>
<td>0.73</td>
</tr>
<tr>
<td>( DIST )</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>5.72</td>
<td>0.60</td>
</tr>
<tr>
<td>( R^2 )</td>
<td>0.05</td>
<td></td>
<td>0.91</td>
<td></td>
<td>0.92</td>
<td></td>
</tr>
<tr>
<td>Adjusted ( R^2 )</td>
<td>0.04</td>
<td></td>
<td>0.90</td>
<td></td>
<td>0.89</td>
<td></td>
</tr>
<tr>
<td>( F ) test statistic</td>
<td>10.39</td>
<td>&lt;0.01</td>
<td>1,038.22</td>
<td>&lt;0.01</td>
<td>464.12</td>
<td>&lt;0.01</td>
</tr>
</tbody>
</table>
2.4 Data transformation

As mentioned in the first session of this chapter, linear regression analysis starts with an assumption that the relationship between $Y$ and its determinants is linear. This does not limit the application of the linear regression method when non-linear relationships are involved. By transforming the original variables, the non-linear relationship can be modelled. This is because the linear regression method requires only ‘linear in parameters’, not ‘linear in variables’. For example, in equation (5) $Y$ and $X$ are not linear in variables (i.e., they are not linearly related). But $\ln(Y)$ and $X$ are linearly related and the corresponding parameter can be estimated by the linear regression method. Hence $Y$ and $X$ are linear in parameters.

$$\ln(Y) = \beta_1 + \beta_1 X + \varepsilon$$ (5)

An examination of the scatter plot between $Y$ and each of the independent variables may be helpful in determining whether transformation is needed. Figure 4.5.4 gives some useful function forms with transformations and examples of their corresponding scatter plots. Note that all patterns except for the first plot show a non-linear relationship between $Y$ and $X$. However, the models for all six cases are in linear form, and can be estimated using the OLS method. When used properly, linear regression models can be used to model a wide range of non-linear relationships. Figure 4.5.4 gives only a fraction of all the possible cases.

It’s worth noting that when coefficients take different values, patterns can look rather different from the ones shown in Figure 4.5.4. Moreover, when $Y$ is affected by more than one independent variable, scatter plots between $Y$ and only one independent variable are unlikely to be informative. In practice, the determination of transformation needed is often an empirical issue. Researchers try several functional forms with different transformations, and choose the one that fits the data the best.

In planning research, natural log transformation and quadratic transformation have the widest application. For example, in Chellman et al. (2011), the school enrolment variable was log
Regression analysis in planning research

transformed. When natural log transformation is used, the transformed variable (dependent or independent) is viewed as a percentage change measurement of the original variable. The interpretation is changed accordingly. Table 4.5.3 gives examples of coefficient interpretation of different types of log transformations.

Figure 4.5.4 Useful transformations.
Natural log transformation is necessary when the variation of $Y$ gets larger as the value of $X$ increases. It is also useful when a variable has a fat right tail in its distribution (e.g., there are some large outliers). Therefore, by constructing scatter plots between $Y$ and each of the independent variables or histograms, one can decide whether a natural log transformation is needed. Natural log transformation is done by using function LN() in Excel. Note that we use only natural log transformation in linear regression analysis. Log transformations using other bases do not give the same coefficient interpretation in Table 4.5.3.

Quadratic transformation is the creation of the second-order term (i.e., the squared values) of a variable. This type of transformation is routinely used to model a non-linear relationship or in trend analysis, which is a standard way of looking at long-term trends in data. A general quadratic model is

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$$

(6)

When $X$ is a time index (e.g., $X=1$ in the first period, and so on), equation (6) is a trend model. By varying the values of $\beta_1$ and $\beta_2$ different types of trend patterns evolve, as shown in Table 4.5.4. This allows a wide range of non-linear patterns to be modelled using linear regression models.

Using the energy performance example again, this application of natural logarithm and quadratic transformations is illustrated here.

A scatter plot between PRICE and EP, and histograms of the two variables are given in Figure 4.5.5. In the scatter plot PRICE initially increases as EP improves, but then decreases after EP goes beyond 94. This is similar to case (e) in Table 4.5.4. A quadratic transformation of EP could be useful in this case. The histogram of PRICE indicates a natural log transformation may be helpful, as there are some outliers present. Based on the analysis of these charts, we can try a natural log transformation on PRICE and quadratic transformation on EP.

For comparison purposes, two regression models were estimated as follows.

Model 4: $\ln(\text{PRICE}) = \beta_0 + \beta_1 \text{EP} + \beta_2 \text{SIZE} + \beta_3 \text{NEW} + \beta_4 \text{FAST} + \beta_5 \text{DIST} + \epsilon$

Model 5: $\ln(\text{PRICE}) = \beta_0 + \beta_1 \text{EP} + \beta_2 \text{EP}^2 + \beta_3 \text{SIZE} + \beta_4 \text{NEW} + \beta_5 \text{FAST} + \beta_6 \text{DIST} + \epsilon$
Model 4 has an $R^2$ of 83.55 per cent. This is a good fit of the data. Nevertheless, it is not appropriate to compare the $R^2$s between Model 3 and Model 4. This is because the dependent variables in the two models are different. Hence it is incorrect to say that Model 4 is outperformed by Model 3 by 9 per cent (i.e., 92 per cent – 83 per cent).
### Table 4.5.5 Regression using LN(PRICE) as a dependent variable

<table>
<thead>
<tr>
<th></th>
<th>Model 4</th>
<th></th>
<th>Model 5</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Coefficients</td>
<td>P-value</td>
<td>Coefficients</td>
<td>P-value</td>
</tr>
<tr>
<td>Intercept</td>
<td>11.5710</td>
<td>0.0000</td>
<td>-0.3975</td>
<td>0.9317</td>
</tr>
<tr>
<td>EP</td>
<td>0.0021</td>
<td>0.5579</td>
<td>0.2652</td>
<td>0.0097</td>
</tr>
<tr>
<td>EP2</td>
<td>-</td>
<td>-</td>
<td>-0.0014</td>
<td>0.0103</td>
</tr>
<tr>
<td>SIZE</td>
<td>0.0096</td>
<td>&lt;0.0001</td>
<td>0.0096</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>NEW</td>
<td>-0.0529</td>
<td>0.3616</td>
<td>-0.0042</td>
<td>0.9444</td>
</tr>
<tr>
<td>FAST</td>
<td>0.0643</td>
<td>0.0567</td>
<td>0.0647</td>
<td>0.0520</td>
</tr>
<tr>
<td>DIST</td>
<td>0.000028</td>
<td>0.3631</td>
<td>0.000038</td>
<td>0.2088</td>
</tr>
<tr>
<td>R Square</td>
<td>0.8355</td>
<td></td>
<td>0.8406</td>
<td></td>
</tr>
<tr>
<td>Adjusted R Square</td>
<td>0.8316</td>
<td></td>
<td>0.8360</td>
<td></td>
</tr>
<tr>
<td>F test statistic</td>
<td>216.38</td>
<td></td>
<td>186.27</td>
<td></td>
</tr>
<tr>
<td>F test P-value</td>
<td>&lt;0.001</td>
<td></td>
<td>&lt;0.001</td>
<td></td>
</tr>
</tbody>
</table>

The coefficient estimates are different in Model 4. This is because the natural log transformation of \(\text{PRICE}\) changes the meaning of these coefficients. For example, the coefficient estimate of \(\text{SIZE}\) is 0.0096 now. It indicates that for each squared metre increase in size, house prices will increase by 0.95 per cent on average. This is a non-linear relationship between \(\text{PRICE}\) and \(\text{SIZE}\), because the same percentage of changes in \(\text{PRICE}\) is different across the value range of \(\text{PRICE}\). More specifically, for larger houses an increase of size by one squared meter will cause house prices to increase more in absolute value (albeit by same percentage). This type relationship is illustrated in case (d) in Figure 4.5.4.

In Model 5, it is assumed that as energy performance score increases the percentage changes in \(\text{PRICE}\) will increase, but at a decreasing rate. Results are given in the last two columns of Table 4.5.5.

Model 5 offers improvement over Model 4, as evident from the increased \(R^2\) and significant \(t\) test results for the second-order term of \(\text{EP}\). The results also support our assumption. House prices appreciate as \(\text{EP}\) increases (i.e., the coefficient estimate of \(\text{EP}\) is positive) but the rate of change slows down (i.e., the coefficient estimate of squared \(\text{EP}\) is negative).

### 2.5 Diagnostics tests

The simplicity of OLS comes with a price. There are some strong assumptions on the error term \(\epsilon\) in OLS estimations. If any of these assumptions is violated, the OLS results could be misleading. The solution is to put the model through a series of tests to check if these assumptions hold. It is also the convention to check if the model suffers from mis-specification (e.g., a non-linear relationship is not accounted for) and multicollinearity (e.g., independent variables are highly correlated) issues. Commonly used diagnostics tests for OLS models are summarized in Table 4.5.6. Most statistical software packages can perform these tests. For statistical properties of the tests Wooldridge (2009) is a good textbook to consult. A good example of applications of these tests can be found in Greasley (2011).
Regression analysis in planning research

3. Practical issues

Planning outcomes can be measured directly or indirectly. For example, a local government introduced a congestion charge in certain parts of the city to reduce traffic and air pollution in the city centre. The effect of this policy can be measured directly by observing the changes in traffic and air quality in affected areas, or indirectly by testing if property prices in the city centre increased after the implementation of the restriction. In the latter case, the assumption is that properties away from the city centre will be sold at a discount to compensate the congestion charge generated by driving to the city centre, all else being equal.

When planning outcomes are measured directly, regression techniques are often used to identify the net policy impact. Considering the congestion charge example again, a researcher can obtain air quality changes after the policy comes into force. However, during the sample period other factors affecting air quality are not likely to remain constant. If their values change as well, the observed air quality changes contain more than policy effects. By regressing the planning outcome variable (i.e., air quality changes) on both the policy factor (e.g., a variable equals one for areas affected by the charges, and zero otherwise) and other control variables (e.g., wind speed, temperature), the effect of congestion charge will be estimated accurately.

When direct measurement is not available, changes in property or land prices are often used as a proxy of planning outcomes. Regression analysis using property or land prices as dependent variables is also called hedonic price modelling (HPM). The theoretical underpinnings of HPM were first established by Rosen (1974), by which the prices of durable goods are associated with economic values of individual attributes. There is a sizable literature of planning studies using HPM. For example, Bartholomew and Ewing (2011) critically evaluated hedonic price studies on the values of pedestrian- and transit-designed development; Machin (2011) surveyed the literature of hedonic valuation of good schooling.

There are several issues facing planning researchers with respect to HPM.

3.1 Omitted variable bias

Properties are complex goods, whose values are determined by a myriad of attributes. If any one of these price determinants is missing from the hedonic price model, planning policy effect estimation could be biased. This is called omitted variable bias. The magnitude and the direction of omitted variable bias depend on the covariance between omitted variables and planning factors included in the model, and the coefficient of omitted variables. Hence it is not straightforward to estimate or adjust omitted variable bias.
When omitted variable bias is suspected, a review of empirical literature is always helpful to identify any important variables missing from the model. If information on one important variable is unavailable in the dataset, the planning factor’s coefficient estimate may still be unbiased as long as the omitted variable is not significantly correlated with the included planning factor. For instance, if $SIZE$ is missing from Model 5, the coefficient estimate of environmental performance rating variables will not be affected if $SIZE$ and $EP$ are not correlated. If, however, there are empirical evidences showing that larger houses tend to have poorer energy performance scores, $EP$ and $EP^2$’s coefficient estimates may be biased, and should be interpreted with caution. The researcher should either collect data for the missing variable or use proxies. If no new information is available, readers should be warned that the regression results may be misleading.

3.2 Mis-specification
Linear relationship is only one of many possible relationships between planning outcomes and their determining factors. For example, it is established that property prices are non-linearly related to size and age. The convention is to include the second-order terms of these variables to capture any non-linear patterns. The relationship regarding other factors, however, is not as well established. When a variable enters the model with a wrong functional form (e.g., a quadratic term is missing), mis-specification bias may present. The HPM will produce unreliable coefficient estimates and test results.

The Ramsey RESET test is commonly used to detect mis-specification problems. If the test results turn out to be significant, the model needs to be adjusted. This is often an empirical issue. Planning researchers need to run several plausible data transformations (e.g., square, logarithms) before the best functional form can be determined.

3.3 Multicollinearity
Housing attributes are often correlated. For example, house size largely determines the number of bedrooms and bathrooms. When the correlation among independent variables is significant (e.g., sufficiently large), both coefficient estimates and hypothesis testing results are adversely affected. More specifically, the regression technique is not capable to separate the effects of correlated variables, and the standard error of coefficient estimates will be inflated. In some extreme cases, a model may have a significant $F$ test result, which means at least one independent variable is useful, but with insignificant $t$ test results for all independent variables. The contradictory findings are a classic sign of multicollinearity problems. When multicollinearity is serious, some coefficient estimates may have signs that are opposite to what is suggested in theories or empirical literature. This will cause confusion and difficulties in results interpretation.

To detect multicollinearity problems, variance inflation factors (VIFs) are often calculated for each independent variable. VIF is a measurement of multiple correlation between an independent variable and all other regressors. The rule of thumb is that if a variable’s VIF is greater than 5 there is potentially a harmful multicollinearity problem.

If a variable is found to be highly correlated to other independent variables, it should be replaced with a different measurement of the same housing attribute, but with lower correlation with the rest of the regressors. If such a variable is not available, the solution is usually to simply drop the problematic variable from the model.
4. Summary

The multiple linear regression method is a useful tool to identify linear relationships between a variable and a group of factors affecting its values. In this chapter we introduced the OLS method for linear regression estimation and its applications in planning research. Some practical issues regarding hedonic price modelling of planning issues were also discussed. The procedures of a standard linear regression analysis are summarized in Figure 4.5.6.

A final note to be made is the analysis of time series data. It is worth noting that the OLS method should be used for cross-sectional data. When applied to time series data, some OLS assumptions, such as zero serial autocorrelation, are often violated. Time series data are often not stationary (i.e., the mean, variance, and covariance are not constant over time). This will give rise to spurious regression results when OLS is used. The rule of thumb is if a model’s R square is greater than its DW test statistic, there are potentially spurious regression problems. Some standard tools, such as the Unit Root test, can be used to check the stationarity of the data. The OLS method can be used if time series are stationary or co-integrated. Planning researchers should be aware of the limitations of linear regression techniques in time series analysis. The test of stationarity and time series analysis techniques are beyond the scope of this chapter. Interested readers can find some less maths-intensive materials on this topic in Wooldridge (2009).

---

Figure 4.5.6 Suggested steps for regression analysis.
### Appendix

Table 4.5.A1  Selected publications in planning research that used regression techniques (2010–2011)

<table>
<thead>
<tr>
<th>Paper</th>
<th>Journal</th>
<th>Country</th>
<th>Sample size</th>
<th>Number of variables in final model</th>
<th>Planning issues studied</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chellman et al. (2011)</td>
<td>Journal of the American Planning Association</td>
<td>USA</td>
<td>5592</td>
<td>16</td>
<td>Subsidized Housing</td>
</tr>
<tr>
<td>Stagoll et al. (2010)</td>
<td>Landscape and Urban Planning</td>
<td>Australia</td>
<td>80</td>
<td>6</td>
<td>Urban Design – Conservation</td>
</tr>
<tr>
<td>Yusof &amp; Shafiei (2011)</td>
<td>Housing Studies</td>
<td>Malaysia</td>
<td>118</td>
<td>4</td>
<td>Urban Design – Innovation</td>
</tr>
</tbody>
</table>
Notes

1 A statistical test always has two hypotheses – a null hypothesis and an alternative hypothesis. These two hypotheses are collectively exhaustive and mutually exclusive. For example, a test has a null hypothesis of $\beta = 5$ and an alternative hypothesis of $\beta \neq 5$. The two hypotheses do not have anything in common (mutually exclusive), but include all possible values for $\beta$ (collectively exhaustive). This ensures a conclusion can be drawn (i.e., whether $\beta$ equals five or not) at the end of test. Therefore it is important to make one of the hypotheses exactly what one wants to test so that the research question can be answered by the end of hypothesis testing.

2 The definition and calculation of $p$-value requires some statistical background. Interested students are advised to read relevant chapters in statistical textbooks. In a nutshell, $p$-value is calculated based on the assumption that the null hypothesis is true. If $p$-value is too small, one can conclude that the assumption is not valid, and consequently reject the null hypothesis.

3 Linear in variables means original variables (without being transformed in any way) are linearly related, and linear in parameters means the transformed variables are linearly related.

4 $R^2$ square measures the percentage of variation in $Y$ captured by all independent variables in a regression model. If the dependent variable is different in two models, their variance (i.e., the variation) will be different too. This makes the comparison of $R^2$ square based on ‘unlevelled ground’.

5 This test is designed to identify any missing non-linear relationship in the model (e.g., a missing second-order term). The null hypothesis in a RESET test is that the model is correctly specified. If the null hypothesis is rejected, it is often helpful to check if any higher-order term or cross-terms (i.e., product of two independent variables) are missing from the model.

References


Helen Bao

